5-8-2019

A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning

Jared Taylor
Iowa State University, jaredtay@iastate.edu

Chien-Ping T. Chou
Iowa State University, cchiou@iastate.edu

Leonard J. Bond
Iowa State University, bondlj@iastate.edu

A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning

Abstract

The ability of terahertz (THz) electromagnetic waves to penetrate a wide range of materials gives potential for diverse applications in nondestructive evaluation, biomed, and agriculture and there has been rapid expanding both in its use. One possible application is in relation to corn breeding, specifically when the doubled haploid method is used as a process that greatly speeds up plant breeding, and this requires seed sorting. Haploid kernels are induced in corn plants in order to decrease the time to reach homozygous genetic corn lines. These haploid kernels must be separated from the surrounding diploid kernels; presently this is labor intensive and performed using visual markers. This current work represents a proof of concept study which sought to determine if haploid classification can be automated using terahertz time domain spectroscopy (THz-TDS) with data analysis paired with a machine learning algorithm, such as a probabilistic neural network (PNN). In this work, a THz-TDS system was used to collect time domain waveforms from a sample of mixed haploid and diploid corn kernels. Effects of variabilities in beam focus and kernel geometry were reduced by taking multiple scans at different heights. The waveform data were then transformed to the frequency domain and further classified by PNN with a training set random subsampling technique. Leave-one-out and K-folds cross-validation procedures were used to train the model. The preliminary results show promise yielding an average classification rate of 75 percent correct by 5-fold cross-validation.

Keywords

Artificial neural networks, Machine learning, Electromagnetism, Horticulture techniques, Terahertz time-domain spectroscopy

Disciplines

Artificial Intelligence and Robotics | Horticulture | Plant Breeding and Genetics

A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning

Cite as: AIP Conference Proceedings **2102**, 080001 (2019); https://doi.org/10.1063/1.5099809
Published Online: 08 May 2019

Jared Taylor, Chien-Ping Chiou, and Leonard J. Bond

A Methodology for Sorting Haploid and Diploid Corn Seed Using Terahertz Time Domain Spectroscopy and Machine Learning

Jared Taylor[1, a)] and Chien-Ping Chiou[1, b)] and Leonard J. Bond[1, c)]

[1]*Center for Nondestructive Evaluation, Iowa State University, Ames, Iowa 50011, USA*

[a)]Corresponding author: jaredtay@iastate.edu
[b)]cchiou@iastate.edu
[c)]bondlj@iastate.edu

Abstract. The ability of terahertz (THz) electromagnetic waves to penetrate a wide range of materials gives potential for diverse applications in nondestructive evaluation, biomed, and agriculture and there has been rapid expanding both in its use. One possible application is in relation to corn breeding, specifically when the doubled haploid method is used as a process that greatly speeds up plant breeding, and this requires seed sorting. Haploid kernels are induced in corn plants in order to decrease the time to reach homozygous genetic corn lines. These haploid kernels must be separated from the surrounding diploid kernels; presently this is labor intensive and performed using visual markers. This current work represents a proof of concept study which sought to determine if haploid classification can be automated using terahertz time domain spectroscopy (THz-TDS) with data analysis paired with a machine learning algorithm, such as a probabilistic neural network (PNN). In this work, a THz-TDS system was used to collect time domain waveforms from a sample of mixed haploid and diploid corn kernels. Effects of variabilities in beam focus and kernel geometry were reduced by taking multiple scans at different heights. The waveform data were then transformed to the frequency domain and further classified by PNN with a training set random subsampling technique. Leave-one-out and K-folds cross-validation procedures were used to train the model. The preliminary results show promise yielding an average classification rate of 75 percent correct by 5-fold cross-validation.

INTRODUCTION

Corn is ubiquitous in this modern world and improved varieties are being sought. One of the first steps in breeding corn is the production of homozygous candidate lines. In a homozygous kernel, both sets of chromosomes are identical. If the chromosomes are identical, the next inbred generation will be determined and all progeny kernels will be identical. Traditional corn breeding involves many generations of inbreeding and a 99% homozygous corn line can be established in typically six to eight generations. Using the doubled haploid breeding technique (DH), the process can be reduced to only two to three generations [1]. The DH method involves induction of haploid kernels in a genetic stock. Haploid kernels, as opposed to diploid kernels, have only half of the genetic material (one set of chromosomes). These haploids can later be treated with a chemical called colchicine to induce duplication of the chromosomes, producing a homozygous diploid plant in just two generations.

Haploid kernels appear naturally, but are very rare. Haploid inducer lines must be used to increase the proportion of haploids that appear on an ear of corn. Boote et al. [2] reports that a hybrid line crossed with an inducer line will produce approximately 10% haploids on the ear of corn. These haploid inducer lines are bred not just to induce haploids, but also to make them distinguishable from the diploids. If the inducer line is successful in merging with the embryo, the kernel will have a purple tinted embryo and endosperm, the marks of a diploid kernel. If the inducer line doesn't merge with the embryo, the endosperm will remain purple, but the embryo will be colorless, a haploid kernel. Kernels pollinated by a third party are identifiable as well, since neither endosperm nor embryo will be purple. Figure 1 shows examples of the difference between haploid and diploid kernels.

45th Annual Review of Progress in Quantitative Nondestructive Evaluation, Volume 38
AIP Conf. Proc. 2102, 080001-1–080001-10; https://doi.org/10.1063/1.5099809
Published by AIP Publishing. 978-0-7354-1832-5/$30.00

FIGURE 1. Visual haploid/diploid kernel discrimination is performed by looking at both the embryo and the endosperm. Both kernels have purple endosperms, the right kernel has purple embryo and left does not. The left kernel is haploid, and that on the right is diploid.

The industry standard method for separating haploid from diploid kernels at present is manual and performed by trained technician. This can be very expensive to sort the numbers of kernels needed. There is therefore a need for automation to be introduced in this industry. Other methods, using various sensor technologies have been explored [2-7], with this work reporting a proof of concept assessment of THz electromagnetic waves, which offers potential for automation.

THz energy was used for kernel interrogation and data collection in this work. This high frequency electromagnetic radiation has great potential because the radiation can penetrate dielectric materials, with a response that is sensitive to composition and internal structure and it gives data for collection and analysis. Such THz technology is relatively new and has exhibited great potential in a number of industries and applications [8-10]. The resulting data are taken in the form of a time-domain-spectra. Each kernel was raster scanned, with typically 51 potentially usable waveforms collected per kernel. The data in each waveform is transformed to the frequency domain using a Fourier transform, and each frequency is treated as a separate dimension in the data. This data set is both large and multivariate, requiring a machine learning (ML) model for analysis. A probabilistic neural network (PNN) was applied to the data and used to classify each waveform, and by proxy each kernel, as either haploid or diploid. Using this approach the classification process can potentially be automated.

Terahertz Nondestructive Evaluation

Interest in the use of sensing in the THz regime has been growing in the NDE community. Stoik et al. [11] examined aircraft composites for defects such as voids, delaminations, mechanical damage, and heat damage. Hsu et al. [12] demonstrate the ability of THz detect defects which were found behind other defects in multi-layer composites. This capability overcomes one significant a limitation for ultrasonic inspection which is due to the "shadow effect". Lopato and Chady [13] did show how THz-NDE can be used to find damage caused by mechanical impact in basalt fiber reinforced composites. Other applications which have included the inspection of coatings [14], turbine blades [15], and even pharmaceuticals [16] and radar-dome inspection [17] have been documented.

Machine Learning

Machine learning is an approach used to investigate data by using computer algorithms that can, with experience, improve in performance, without being explicitly programmed so to do. As a form of artificial intelligence, this approach borrows heavily from such disciplines as statistics and optimization. Large or complex data sets can be analyzed for purposes such as regression, classification, or clustering. Successful examples of applications include:

facial recognition, spam email filtering and optical character recognition. Machine learning is used extensively in the field of chemometrics when experimental data sets are large and complicated, and an inversion must be performed to correlate and determine experimental conditions.

Each model has parameters that define its behavior. These parameters must first be optimized to improve the performance of the model on unknown (or test) data. Cross Validation (CV) is the system for evaluating a machine learning model using known data. Two CV methods were used in this work. The first, called leave-one-out (LOO) works by setting aside a single data point, training with the others, and testing the model using the data point set aside previously. This method is very simple, but expensive in practice, growing more costly in terms of time and computational resources as the size of the training set increases. It is the same approach as the statistical resampling tool called the jackknife. The second CV tool used is called K-folds, where the training set is divided into K groups. It follows the same steps as LOO, except it is applied to the set of folds, or groups, instead of the raw data. Each group is set aside in turn for testing, while all others are used for training. With these two methods, the performance of a machine learning model when introduced to new data can be evaluated using only information previously known [18].

METHODS

The THz-TDS system used in this work for data collection is called a "TPI Imaga," manufactured by Teraview. It uses a mode-locked 100 femtosecond Ti:Sapphire laser to drive a photoconductive antenna using a lock-in amplifier. Detection is performed using a similar photoconductive antenna. The beam is highly focused at 50 mm focal length with a beam-width at full-width half-maximum of 0.8 mm. The transmitter and receiver are oriented for reflection mode at 17° from the normal. The sensor head is mounted on a 3D gantry capable of 2D translation and raster scanning. An ad-hoc plastic tent was used to provide a controlled atmosphere and to isolate the air around the beam. Dry air was used and it was pumped into the tent to limit water vapor related interference effects. The THz-TDS system is shown In Fig 2.

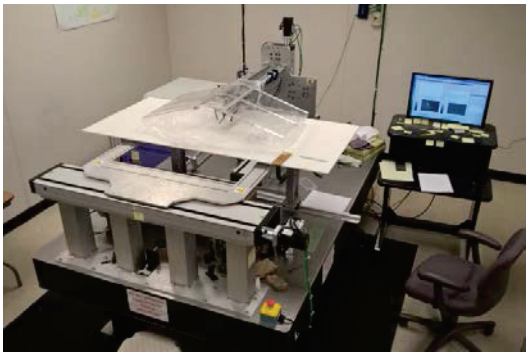

FIGURE 2. THz-TDS system in the THz lab.

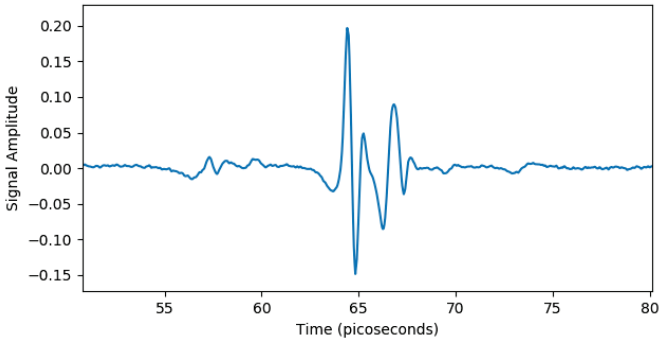

FIGURE 3. Example of a waveform taken for the reflection from the surface of a kernel. Because of the geometry and random nature of the corn kernel, the waveform can be complex.

Sample Corn Kernels

The sample set used consists of 91 hand-picked corn kernels. They were carefully selected with two goals in mind, first, they had to be readily classifiable haploid or diploid by visual inspection, and second, they had to have a relatively flat front face so as to enable them to lay parallel with the translation plane. This geometry was to establish favorable conditions for scanning the embryo. It is understood that in any practical application there will be more varied orientation and these conditions will not be constantly met.

The kernels were laid out in a grid, sufficiently spaced so that the signal would go to zero between kernels. This helped in image segmentation and for data labeling. In a future implementation this orderly layout could potentially be arranged mechanically. The corn kernels used in the analysis are shown in Fig. 4. The samples were categorized into haploid and diploid using standard visual markers. These identification data form the labels that were used for training the PNN. The data set was built by collecting a time-domain waveform (A-scan) from each point in a 2D matrix covering the sample with resolution of 0.5 mm x 0.5 mm. Data for each pixel in the scan represents the max

amplitude of an A-scan taken at that location, with such images forming parts a to d in figure 6. Each kernel had on average 215 A-scans associated with it. Table 1 shows the details of the data set parameters.

FIGURE 4. Corn kernel sample used in the analysis.

The time gate used for capturing the waveforms was long compared to the length of the data waveform needed to perform the analysis. The gate length was set so as to accommodate the varying geometry of the corn kernels, and enable easier post processing by including reference points in the recorded data. The gate had to be long enough to enable data to be measured for the tallest corn kernel and the reflection of the signal from the baseplate when the scanner was between kernels. This requirement guided selection of a 157 picosecond time gate. This time gate was sampled 4096 times, resulting in a sampling time step of 38.3 femtoseconds, or a sampling frequency of about 26 THz.

TABLE 1. Data collection parameters.

Scan Parameter	Value
Scan Resolution	0.5 mm x 0.5 mm
Number of Scans	4
Scan Height	4, 6, 7, 8 mm from baseplate
Total A-scans on Kernel Faces	19,533
A-scans Per Kernel	215
Time Resolution	38.3 femtoseconds
Total Kernels, Diploid, Haploid	91, 44, 47

Focus Compensation

The complicated shape of the corn kernel presents a significant technical challenge in this work. The kernel shape and thickness/height can vary drastically and randomly from kernel to kernel. The tallest kernel had a height of 7.49 mm, while the smallest kernel is only 3.87 mm tall. To mitigate this problem, four scans were performed with THz field focus set at 4 mm, 6 mm, 7 mm, and 8 mm heights, measured from the sample plate as shown in Fig 5. This produced four data sets. The four were reduced to one by comparing peak-to-peak amplitude of the four choices at each pixel and saving the waveform with largest peak-to-peak amplitude. This process with various scan heights and resulting data images are shown in Fig. 6.

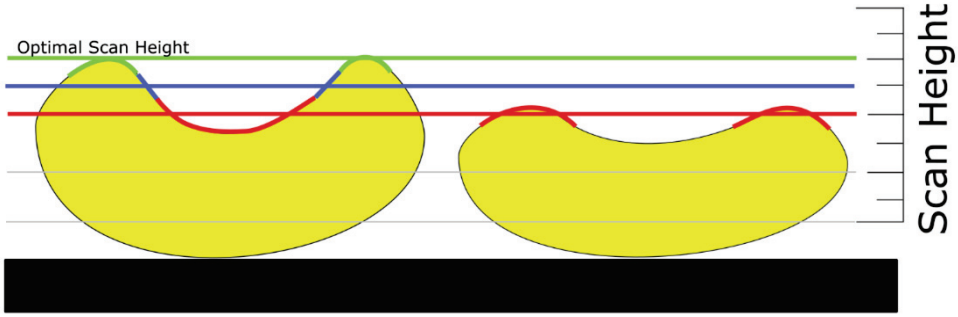

FIGURE 5. Different scan heights allow for different kernel surfaces to be in focus.

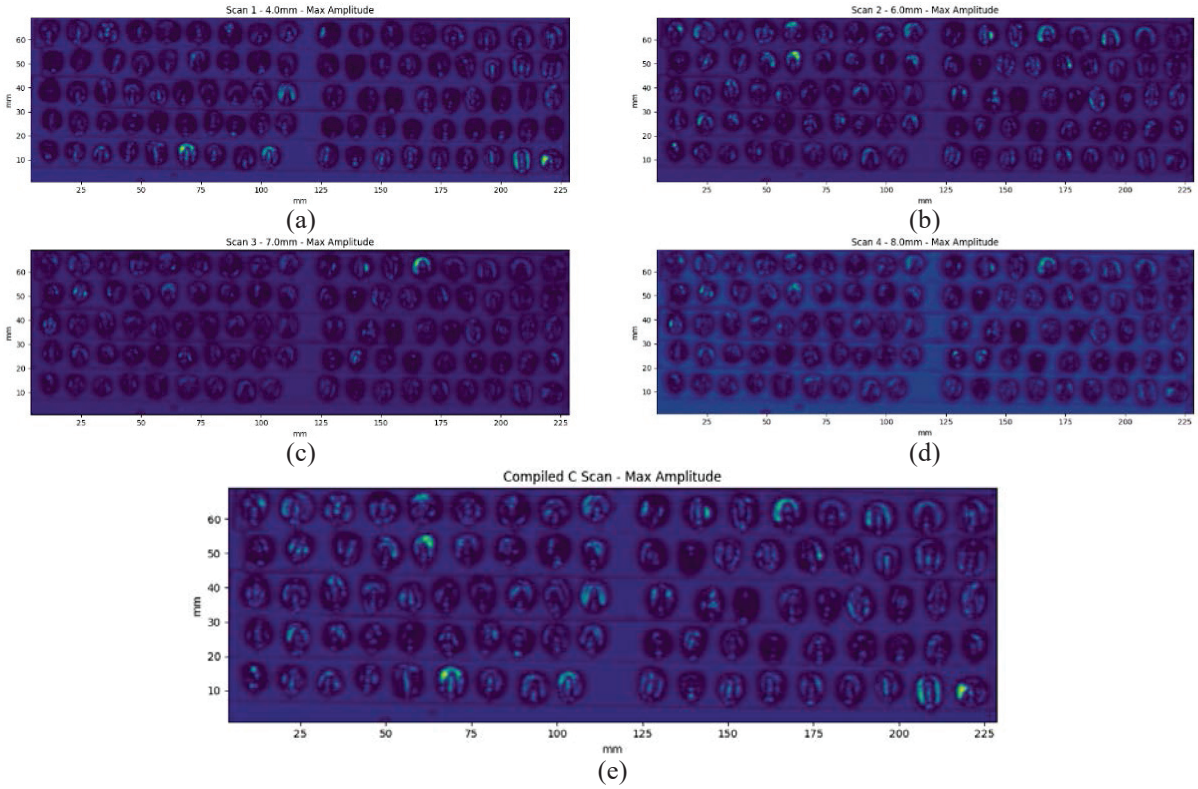

FIGURE 6. Scan images at various focal plane levels (a-d) and compound image (e) max amplitude in the time domain.

Probabilistic Neural Networks

The probabilistic neural network (PNN) was introduced by Specht [19]. It is closely related to what Mitchell calls "instance based learning" using Gaussian radial basis functions [20]. This scheme is a modified artificial neural network that can map any input to any number of output discrete classifications. The underlying equation uses a probability density function developed by Parzen [21] to calculate decision boundaries. These boundaries can be linear or non-linear, and they are selected depending on the data set. As an approach a PNN stands at the edge between more traditional artificial neural networks called "eager learners", with extensive training stages up front, and so called "lazy", methods called where training doesn't happen until a new data point is presented.

Parzen [21] presented a class of PDF estimators approach the parent density function so long as it is continuous. This estimation of the probability allows the use of a Bayes rule, and thereby creation of a Bayes learner. This work is a binary decision ($d(X)$) made by way of PNN, having the rule

$$d(X) = \theta_A \text{ if } h_A l_A f_A(X) > h_B l_B f_B(X) \textbf{Error! Bookmark not defined.} \quad (1)$$

$$d(X) = \theta_B \text{ if } h_A l_A f_A(X) < h_B l_B f_B(X) \tag{2}$$

where θ_A and θ_B are the output classification decision of either class A or class B; $f(X)$ is the probability density function of the classes; l is the loss function associated with the classes; and h is the a priori probability of the classes. A decision boundary can be drawn according to equation 3.

$$f_A(X) = K\, f_B(X),\ K = \frac{h_B l_B}{h_A l_A} \tag{3}$$

A Bayes decision rule developed in this way will asymptotically approach the Bayes optimal classifier with increasing sample size [19].

The model consists of four layers: the input layer, the pattern layer, the summation layer, and the output layer. The model used in this work uses a summation of spherical Gaussian basis functions centered at each training data point to approximate the underlying probability density function (PDF). The standard deviation σ (what Specht calls the smoothing parameter) of the Gaussians and the cost/prior parameter K must be optimized in order to train the PNN; making this an eager learning model. The following equation is the PDF estimator kernel used in this work.

$$f_p(X) = \frac{1}{(2\,\pi)^{\frac{v}{2}}\sigma^v}\frac{1}{N_\mathrm{p}}\sum_i^{N_\mathrm{p}} e^{\left[-\frac{\left(X-X_{pi}\right)^t(X-X_{pi})}{2\sigma^2}\right]} \tag{4}$$

In equation 4, f is the estimate of the underlying density function for the pth class; v is the number of variables; σ is the smoothing parameter; N is the number of training data points in the pth class; X is the test data point; and X_{pi} is the training data point. Equation 5 is the theoretical equation for the PNN. In practice, not all terms must be present for the performance to be optimal. In practice removing some terms will lighten the computational load. In the final analysis, all that is needed is a comparison between the probability density functions for each class, and like terms can be canceled out. Equation 6 shows the relationship used in this work to apply the PNN.

$$f_p(X) = \frac{1}{N_p}\sum_i^{N_\mathrm{p}} e^{\left[\frac{\left(X-X_{pi}\right)^t(X-X_{pi})}{2\sigma^2}\right]} \tag{5}$$

The structure of the PNN is given in figure 7. It shows a single test data point, two classes and eight data points in the training set. Each data point in the test set must be mapped to each data point in the training set using equation 5 and summed according to the classes in the training set. The output is an estimation of the probability that the test point is in each class. The class with highest probability is chosen as the estimated class of the test data point.

An advantage of the PNN is it can be trained very quickly. Only two parameters must be optimized for performance, after that, all the training data points must be kept on hand. The training process involves applying the PDF estimator kernel equation above for each class; making this, in a way, a lazy learning model as the training is done just before classification. Highly multivariate data is easily handled. It is robust against bad or noisy data so long as the data set is large. New data can be added quickly with or without retraining.

Cross Validation

Validation was performed in two ways in this work, using leave-one-out cross-validation (LOO), and K-folds cross-validation. LOO involves looping across all the data, setting aside one data point at a time, training on all the others, and then testing with the one left out. Spectra from the same kernel can't be used in both the training and testing sets while maintaining the validity of the model. For this reason, spectra from an entire seed were set aside at one time, thus speeding up the validation process. LOO is the simplest form of cross-validation. It gives the most optimistic estimate of the true performance of the model. LOO was used to optimize the σ and cost/prior weights of the model.

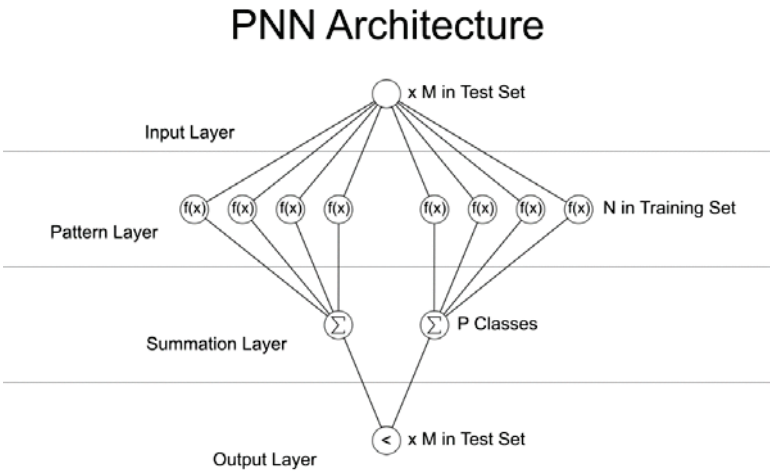

FIGURE 7. PNN model flow chart.

K-folds is a more trustworthy measure of real-life performance. It involves separating the data into K groups, setting a group to be the test group and training on the rest, and then rotating roles across the K groups. For each iteration 1/K of the data set is set aside at random to be the test set, while the rest make up the training set. This must be done many times and the groups must be chosen randomly. In this work K-folds cross-validation was used as a confirmation tool to ensure the LOO cross-validation worked properly, and to give an estimate for the robustness of the method.

RESULTS AND DISCUSSION

The PNN was applied to the data set on a per spectrum basis. Each spectrum was identified according to the class that had the highest probability, according to the PNN. Each seed was classified according to the majority class to which its constituent spectra were assigned. Two PNN models were built to differentiate haploid and diploid corn kernels using the data from the THz-TDS scan. One uses a σ smoothing value of 4.1238e-5 on spectra limited to between 0.0 and 1.0 THz, and the other uses a σ smoothing value of 1.3665e-4 on spectra limited to between 0.0 and 0.5 THz. Extensive cross-validation was performed on both models including LOO cross-validation to evaluate accuracy, and K-Folds cross-validation to evaluate robustness.

Frequency Band Selection

Using the full bandwidth provided by the waveform collected is not valid. Such a spectrum extends to a frequency as high as 12 THz. It is however known that the photoconductive antennas used in this system only have a bandwidth up to 4.0 THz. A model was built to use just this part of the spectrum, but it failed to identify an effective smoothing factor. Similar models were tested all starting from 0.0 THz (dc) and extending to 3.5, 3.0, 2.5, 2.0, 1.5, 1.0, and 0.5 THz. Only the models built with 1.0 THz and 0.5 THz bandwidths showed promise for classification. These values of σ can be used to further train the model by other means, such as the subsampling technique. Table 2 shows the results of the leave-one-out cross-validation.

TABLE 2. Leave-one-out cross-validation results before training set reduction, with 0.5 and 1.0 THz bandwidths.

Performance Marker	0-0.5 THz	0-1.0 THz
Smoothing Factor	1.3665e-4	4.1238e-5
Spectrum Accuracy	52.24 (%)	51.4 (%)
Haploid Accuracy	55.32 (%)	55.3 (%)
Diploid Accuracy	61.36 (%)	56.8 (%)
Kernel Accuracy	58.24 (%)	56 (%)

Training Set Optimization

Both models were optimized by reducing the training set size and randomly choosing which training data points produced the best results [22], measured in terms of correct classification based on optical data. This method was very successful using LOO cross-validation. Table 3 shows the results. As high as a 31.9% improvement in correct kernel classification rate was observed. For the 0.5 THz bandwidth case the training set was reduced to 48.5% of its original size. This result was attained by running the optimizer for 12,656 iterations over about three days using a personal computer. Since it is a purely random phenomenon, there is no guarantee that a global optimum has been reached. The optimizer was stopped when it failed to improve the model after iterating for a long period of time, approximately 12 hours. A possible improvement to this optimizer would be to implement some evolutionary optimization methods, making use of a population of candidates and using crosses and random mutation to vary the training set.

TABLE 3. Leave-one-out cross-validation results after reducing the training set, with 0.5 and 1.0 THz bandwidths.

Performance Marker	0-0.5 THz	0-1.0 THz
Smoothing Factor	1.3665e-4	4.1238e-5
Best Spectrum Accuracy	61.5 (%) (+9.26)	59.6 (%) (+8.2)
Best Haploid Accuracy	87.2 (%) (+31.88)	85.1 (%) (+29.8)
Best Diploid Accuracy	86.4 (%) (+25.04)	90.9 (%) (+34.1)
Best Kernel Accuracy	86.8 (%) (+28.56)	87.9 (%) (+31.9)

Classification Robustness

K-folds cross-validation was used for both models. Good performance in K-folds cross-validation with few folds is a sign of robust models. Both models exhibit good performance in K-folds using 5 folds. Table 4 shows some statistics of the procedure using 5 folds, and data are also given in figures 8 and 9.

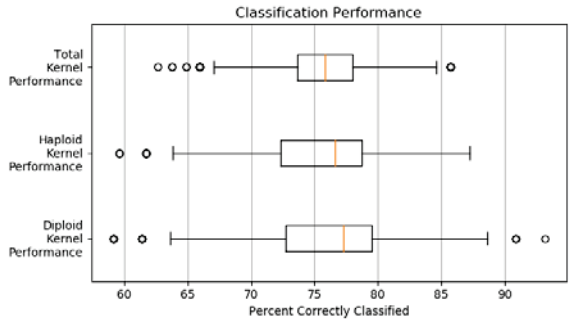

FIGURE 8. K-folds cross-validation results with 5 folds, using the 0.0-0.5 THz band.

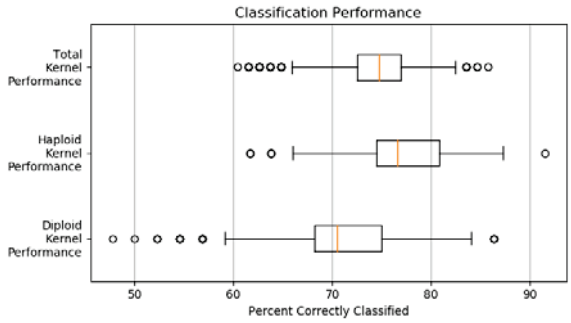

FIGURE 9. K-folds cross-validation results with 5 folds, using the 0.0-1.0 THz band.

TABLE 4. 5-Fold cross-validation results using the subsampled training data set with bandwidths of 0.5 and 1.0 THz.

5-Folds		Spectrum (%)	Kernel (%)	Haploid (%)	Diploid (%)
0.0-0.5 THz	Median	59.29	75.8	76.6	77.3
	S. D.	0.84	3.56	4.67	5.05
N = 2448	Variance	0.70	12.67	21.81	25.5
0.0-1.0 THz	Median	57.24	74.72	76.6	70.45
	S. D.	0.73	3.65	3.98	5.69
N = 1632	Variance	0.53	13.32	15.84	32.38

K-folds cross-validation was performed with 7, 10, 13, 15, and 20 folds. Each increase in number of folds means that smaller group sizes are employed, so as to better approximating the LOO cross validation. K-fold cross validation at 91 folds is equivalent to LOO in this data set. As expected, the performance increases steadily as the

number of folds increases, simulating an increase in training data set size. Figure 10 shows how the performance increases as number of folds increases.

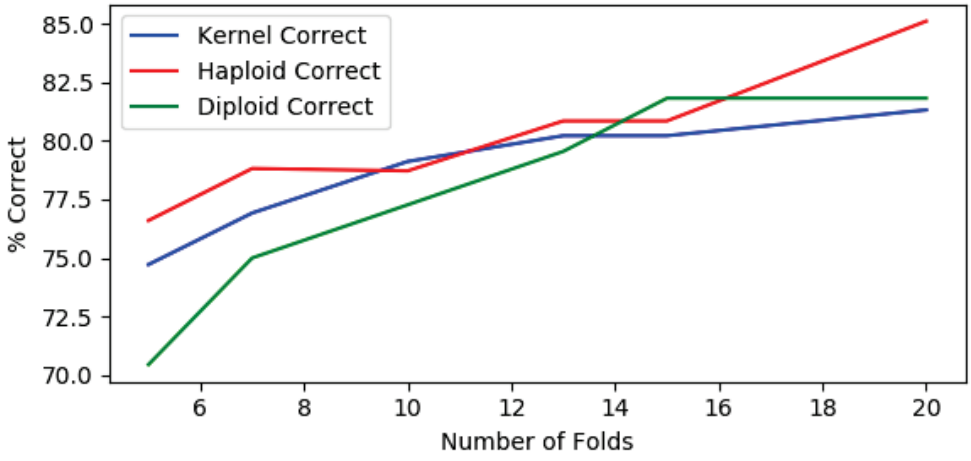

FIGURE 10. Performance progression with increasing number of folds in K-folds cross-validation.

It should be noted that the spectrum classification was relatively poor. Only 59% of the spectra were classified correctly in the 0.5 THz bandwidth case. Because each kernel is identified according to the class that more than 50% of its constituent spectra are identified, the relatively low classification accuracy of a single spectrum can be combined with other spectra on the same kernel. In this way the model can perform much better on each kernel because each kernel has many spectra. A higher spectrum performance would lead to more robust models and would require less data per kernel in the training set.

CONCLUSION

This work represents a novel approach to haploid and diploid classification. A Terahertz time-domain spectroscopy (THz-TDS) system was used to collect many waveforms for a range of corn kernel. The data was transformed to frequency domain then analyzed using a probabilistic neural network (PNN). The PNN classified each spectrum and by proxy each kernel as either haploid or diploid. Leave-one-out cross-validation showed results in as high as 87.9% for correct corn kernel classification, when compared with visual classification. For the same samples the 5-fold cross-validation showed performance averaging approximately 75% correct classification, evidencing some stability in the model. These two results serve as a proof of principle that there may be an application for THz-TDS in classifying haploid and diploid corn kernels.

ACKNOWLEDGEMENTS

The research is supported by the Center for Nondestructive Evaluation, Iowa State University.

REFERENCES

1. B. M. Prasanna, “Doubled haploid (dh) technology in maize breeding: An overview,” in *Doubled Haploid Technology in Maize Breeding: Theory and Practice*, edited by B. M. Prasanna,Vijay Chaikam,George Mahuku (International Maize and Wheat Improvement Center, 2012), pp. 1-8.
2. Brett W. Boote and Daniel J. Freppon and Gerald N. De La Fuente and Thomas Lübberstedt and Basil J. Nikolau and Emily A. Smith, “Haploid differentiation in maize kernels based on fluorescence imaging,” *Plant Breeding,* (2016), **135(4)**:439-445.
3. Roger W. Jones and Tonu Reinot and Ursula K. Frei and Yichia Tseng and Thomas Lübberstedt and John F. McClelland, “Selection of Haploid Maize Kernels from Hybrid Kernels for Plant Breeding Using Near-Infrared Spectroscopy and SIMCA Analysis,” *Applied Spectroscopy*, (2012), **66(4)**:447-450.

4. Gerald N. De La Fuente and Jens Michael Carstensen and Michael A. Edberg and Thomas Lübberstedt, "Discrimination of Haploid and Diploid Maize Kernels via Multispectral Imaging," *Plant Breeding*, (2017), **136(1)**:50-60.
5. Andrew Smelser and Michael Blanco and Thomas Lübberstedt and Axel Schechert and Adam Vanous and Candice Gardner, "Weighing in on a method to discriminate maize haploid from hybrid seed," *Plant Breeding*, (2015), **134(3)**:283-285.
6. Albrecht E. Melchinger and Wolfgang Schipprack and Tobias Wuerschum and Shaojiang Chen and Frank Technow, "Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize," *Scientific Reports*, (2013), 3(2129).
7. Hongzhi Wang and Jin Liu and Xiaoping Xu and Qingming Huang and Shanshan Chen and Peiqiang Yang and Shaojiang Chen and Yiqiao Song, "Fully-automated high-throughput NMR system for screening of haploid kernels of maize (corn) by measurement of oil content," *PLOS ONE*, (2016),**11(7)**:1-14.
8. C. Wang and J. Y. Qin and W. D. Xu and M. Chen and L. J. Xie and Y. B. Ying, "Terahertz imaging applications in agriculture and food engineering: A review," *Transactions of the ASABE*, (2018), **61(2)**:411-424.
9. Taylor, Zachary D and Singh, Rahul S and Bennett, David B and Tewari, Priyamvada and Kealey, Colin P and Bajwa, Neha and Culjat, Martin O and Hubschman, Jean-pierre and Brown, Elliott R and Grundfest, Warren S and Lee, Hua, "THz medical imaging : in vivo hydration sensing," *IEEE Transactions on Terahertz Science and Technology*, (2011), **1(1)**:201-219.
10. J. Suen, "Terabit-per-second satellite links: a path toward ubiquitous terahertz communication," *Journal of Infrared, Millimeter, and Terahertz Waves*, (2016), **37(7)**:615-639.
11. C. Stoik and M. J. Bohn and J. L. Blackshire, "Nondestructive Evaluation of Aircraft Composites Using Transmissive Terahertz Time Domain Spectroscopy," *Optics Express*, (2008), **16(21)**:17039,17051.
12. D. K. Hsu and K. H. Im and C. P. Chiou and D. J. Barnard, "An exploration of the utilities of terahertz waves for the NDE of composites," in Thompson, D. O. and Chimenti, D. E., editors, Review of Progress in Quantitative Nondestructive Evaluation, volume 30, American Institute of Physics (AIP), Conference Proceedings #1335, pp 533-540
13. P. Lopato and T. Chady, "Terahertz detection and identification of defects in layered polymer composites and composite coatings," *Nondestructive Testing and Evaluation*, (2013), **28(1)**:28-43.
14. I. Catapano and F. Soldovieri, "A data processing chain for terahertz imaging and its use in artwork diagnostics," *Journal of Infrared, Millimeter, and Terahertz Waves*, (2017), **38(10)**:1264-1277.
15. T. Fukuchi and T. Ozeki and M. Okada and T. Fujii, "Nondestructive Inspection of Thermal Barrier Coating of Gas Turbine High Temperture Components," *IEEJ Transactions on Electrical and Electronic Engineering*, (2016), **11(4)**:391-400.
16. S. Zhong, "Progress in terahertz nondestructive testing: A review," *Frontiers of Mechanical Engineering* (2018) https://doi.org/10.1007/s11465-018-0495-9.
17. R. Panwar, "Performance and non-destructive evaluation methods of airborne radome and stealth structures," *Measurement Science and Technology*, (2018), **29(6)**:062001.
18. K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*, (CRC Press, 2009).
19. D. F. Specht, "Probabilistic neural networks," *Neural Networks*, (1990), **3(1)**:109-118.
20. T. M. Mitchell, *Machine Learning*, (McGraw-Hill, 1997).
21. E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, (1962), **33(3)**:1065-1076.
22. B. Bolat and T. Yildirim, "Performance increasing methods for probabilistic neural networks," *Information Technology Journal*, (2003), **2(3)**:250-255.